Signal Processing 109 (2015) 226–235

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Robust, transparent and high-capacity audio watermarking in DCT domain

Hwai-Tsu Hu [a],*, Ling-Yuan Hsu [b]

[a] *Department of Electronic Engineering, National I-Lan University, Yi-Lan 26041, Taiwan, ROC*
[b] *Department of Information Management, St. Mary's Junior College of Medicine, Nursing and Management, Yi-Lan 26644, Taiwan, ROC*

ARTICLE INFO

Article history:
Received 24 July 2014
Received in revised form
12 November 2014
Accepted 17 November 2014
Available online 25 November 2014

Keywords:
Blind audio watermarking
DCT-based scheme
Perceptual QIM
Energy variation compensation
Smooth transition

ABSTRACT

This paper presents a novel scheme capable of achieving robust, transparent and high-capacity blind audio watermarking. The watermark embedding is performed by modulating the vectors in the DCT domain subject to an auditory masking constraint. An algorithm has been developed to maintain the energy balance in a frequency band, thus allowing the retrieval of embedded information directly from the designated band. As the proposed watermarking scheme is implemented on a frame basis, the abrupt artefacts in frame boundaries are further rectified via linear interpolation over transition areas. The embedded watermark is shown to be imperceptible to human ears and yet robust against common digital signal processing attacks. The resulting payload capacity is as high as 848 bps. Moreover, not only a 100% recovery of the watermark is guaranteed for non-attack situations but the survival rate is substantially improved in the case of lowpass filtering even at a cutoff frequency as low as 500 Hz.

1. Introduction

Nowadays multimedia data (such as audio, image, and video) are normally stored in digital form, which can be replicated and modified by general users. The widespread use of Internet and wireless networking has made the distribution of multimedia data much easier than ever before. The tendency is further accelerated by a proliferation of smart phones and portable devices in recent years. People around the world keep creating and spreading much multimedia data each day. Unfortunately, the illegal use of multimedia data is also rampant in the digital age. Protection against intellectual property infringement increasingly becomes an important issue. Digital watermarking technology has been considered a promising means to resolve this issue. It is a technique of hiding proprietary information in multimedia data and later extracting such information for purposes of copyright protection, content authentication, ownership verification, etc.

For audio data, watermarking can be implemented in either the time domain [1–3] or transform domains such as the discrete Fourier transform (DFT) [4–6], discrete cosine transform (DCT) [7–10], discrete wavelet transform (DWT) [8,11–14], cepstrum [15–17], singular value decomposition (SVD) [10,18–20]. Transform-domain techniques are generally more robust because they can take advantage of signal characteristics and auditory properties [21]. Wang and Zhao [8], aside from the employment of quantization index modulation (QIM) [22], explored the multiresolution analysis of the DWT and the energy-compression characteristics of the DCT to attain high-performance audio watermarking. Wang et al. [23] later incorporated the well-trained support vector regression into the foregoing DWT–DCT structure where the quantization steps were particularly designed to adapt to the audio signal. The DWT–DCT-based scheme has been reported to achieve effective audio watermarking in robustness, imperceptivity and capacity. However, this scheme suffered from a

* Corresponding author. Tel.: +886 3 9317343; fax: +886 3 9369507.
E-mail address: hthu@mail.niu.edu.tw (H.-T. Hu).

http://dx.doi.org/10.1016/j.sigpro.2014.11.011
0165-1684/© 2014 Elsevier B.V. All rights reserved.

minor shortcoming that there is no guarantee of 100% watermark recovery when the intended payload capacity becomes very high.

To establish a pellucid link between transform domains and human auditory properties, Hu et al. [24] replaced the DWT with the discrete wavelet packet transform (DWPT) that decomposed the audio signal into critical bands. An iterative algorithm was introduced to perform variable-capacity blind audio watermarking subject to a perceptual constraint derived from auditory masking thresholds in different critical bands. The main task was to maintain the spectral flatness and energy level while embedding binary bits in each critical band. Knowing that the computational complexity mostly lies in the preservation of the spectral flatness, Hu et al. [14] instead sought the perceptually tolerable distortion limit for specific subbands and then modified the vector norm collected from the DWT coefficients accordingly. The watermarking schemes in [24,14] have both been shown to be very robust against common attacks and transparent to human ears.

One common feature in the aforementioned schemes is that they all carried out the watermark embedding by modulating low-frequency components. The most commonly used techniques in spectral analysis include DFT, DCT and DWT. The DFT and DCT perform similar functions, which convert a finite list of samples into a combination of basic functions ordered by their frequencies. The difference between these two transforms is that the DFT uses a set of harmonically-related complex sinusoids, while the DCT uses only cosine functions. Because the DCT assumes a continuous symmetry at the boundaries, a signal's DCT representation tends to have its energy compacted in low-frequency coefficients when compared to the DFT. Such a "spectral compaction" property makes the DCT quite suitable for compression applications. The DWT is another popular technique for multiresolution analysis in both time and frequency. However, the required computation is rather complex in comparison with that of the DCT.

It is deduced from the above discussion that the DCT has the merits of high energy compaction in low frequency components and less computational burden. Since the QIM can be easily implemented in the DCT domain, we believe that additional benefits are procurable by directly manipulating the DCT coefficients. Our aim in this study is therefore to develop a DCT-based scheme capable of achieving robust, transparent and high-capacity blind audio watermarking. In the rest of this paper, Section 2 discusses the proposed watermarking scheme in detail. This section has been divided into important subjects including the auditory masking, adaptive QIM, energy compensation and transition smoothing over frames. The focus of this section is on the adjustment of vector norms subject to perceptual considerations. Section 3 presents the performance evaluation in comparison with other recently developed schemes. Finally, Section 4 draws up concluding remarks.

2. DCT-based watermarking

The proposed watermarking scheme is performed on a frame-by-frame basis, where the audio signal is partitioned into frames of length l_f. Each frame contains two sections: a small section of l_t samples is reserved to smooth the transition across frames and the remaining l_w samples are converted to the DCT coefficients, termed c_k's, for watermark embedding. In this study l_t and l_w are tentatively chosen as 64 and 4096, respectively. The adopted frame length l_f is roughly 3.6 times of that used in the MPEG-3 standard. The use of a larger frame will result in a coarse temporal resolution, which is coupled with a high spectral resolution. For a signal sampled at 44.1 kHz, the overall 4160 samples can be translated into a time resolution of 94.3 ms. Moreover, the 4096-point DCT provides a frequency resolution of 5.38 Hz.

Though a large frame is usually unfavorable to the spectral and temporal analysis of transient sounds, this deficiency does not cause any obvious trouble in our watermarking process. As will be clear soon in Section 2.2, except for the special case of setting zero, the phase information characterized by the signs of DCT coefficients is well preserved during watermark embedding. The phase preservation helps to retain the transient behavior of the audio signal, thus averting from the disadvantage of using a large frame. The main concern in this study turns out to be how to manipulate the DCT coefficients in each frame to achieve efficient watermarking.

It has been demonstrated in [14] that a robust and transparent audio watermarking can be achieved by exploiting auditory masking properties [25,26], which suggests that the watermark can be hidden in a sound of higher intensity within each critical band. Because the energy of commonly encountered audio signals is normally centered at the frequency region below 1 kHz, our watermarking process is therefore focused on the first 168 DCT coefficients. These 168 DCT coefficients are further divided into three frequency bands, each covering a range of 301.46 Hz $(=56/l_w \times f_s/2)$ for a sampling rate of $f_s = 44.1$ kHz. With such an arrangement, the estimated auditory masking threshold can be assumed to approximately remain at a fixed level, while each band contains enough DCT coefficients to accommodate the energy variation due to watermarking.

2.1. Auditory masking

According to auditory masking theory [27,28], the inserted watermark will be inaudible if the energy distortion falls below the masking threshold. Hence we take the middle of a frequency band as the representative frequency termed f_{rep} and convert it to a Bark scale via

$$z_{rep} = 13\tan^{-1}(0.00076 f_{rep}) + 3.5\tan^{-1}((f_{rep}/7500)^2). \tag{1}$$

The auditory masking threshold for a band with a center Bark frequency z_{rep} can be estimated using

$$a(z_{rep}) = \lambda a_{tmn}(z_{rep}) + (1-\lambda) a_{nmn}(z_{rep}) \quad [\mathrm{dB}], \tag{2}$$

where λ denotes the tonality factor varying between 0 and 1, $a_{tmn}(z)$ is the tone-masking noise index estimated as $a_{tmn}(z) = -0.275z - 15.025$, and $a_{nmn}(z)$ is the noise-masking noise index usually fixed as $a_{nmn}(z) = -9$ [29]. Since $a(z) \geq a_{tmn}(z)$ no matter what λ is, we can regard

$a_{tmn}(z_{rep})$ as the maximum allowable level of energy variation without causing noticeable distortion, especially at frequencies near z_{rep}. Consequently, the embedded watermark will become imperceptible if the energy variation does not exceed E_{mask}, which is defined as

$$E_{mask} = 10^{\frac{a_{tmn}(z_{rep})}{10}} \times E_c; \quad E_c = \sum_{i=1}^{56} c_i^2. \tag{3}$$

2.2. Norm-space QIM

In our formulation, the 56 coefficients in a frequency band are first categorized into two groups, namely G_1 and G_2, containing the indexes of L_{G_1} and L_{G_2} coefficients respectively. The DCT coefficients in G_1 are intended for watermarking while those in G_2 are for the purpose of absorbing the energy variation. As will be further clarified, a constant energy will allow us to recover the quantization steps from the watermarked signal. Basically, increasing the proportion of L_{G_1} to L_{G_2} will enlarge the payload capacity. However, the amount of energy absorption for each coefficient is limited. It will be difficult to achieve the goal of energy balance if L_{G_2} is too small. Eventually, we choose L_{G_1} and L_{G_2} respectively as 48 and 8 through experimental exploration.

The indexes in G_2 can be selected randomly. In this study a pseudorandom generator is employed to produce a sequence of 56 random numbers. After sorting the random numbers, we record the indices of the first 8 elements as the group G_2 for energy adjustment. The remaining 48 indexes belong to the group G_1. The seed used to generate the random sequence is regarded as a secret key. To embed a binary bit w_b into a selected c_k from G_1, we utilize the QIM rule [22] such that

$$\check{c}_k = \begin{cases} \left\lfloor \frac{c_k}{\Delta} + 0.5 \right\rfloor \Delta, & \text{if} \quad w_b = 0; \\ \left\lfloor \frac{c_k}{\Delta} \right\rfloor \Delta + \frac{\Delta}{2}, & \text{if} \quad w_b = 1, \end{cases} \tag{4}$$

where $\check{c}_k$ denotes the quantized version of c_k. $\lfloor \bullet \rfloor$ stands for the floor function. In such a manner, a total amount of L_{G_1} bits is embedded in the frequency band.

To improve the watermark robustness, we resort to the strategy presented in [13,14], where the QIM is carried out in the norm space. Without loss of generality, we assume that each vector $\mathbf{v}_k$ is constituted by l_v DCT coefficients and L_{G_1} is divisible by l_v, i.e.,

$$\mathbf{v}_k = \left[c_{k_1} c_{k_2} \cdots c_{k_{l_v}} \right], \quad k = 1, \ldots, L_{G_1}/l_v; \quad k_i \in \{1,\ 2,\ \ldots,\ 56\}. \tag{5}$$

The vector norm of $\mathbf{v}_k$, termed σ_k, is simply computed as

$$\sigma_k = \|\mathbf{v}_k\| = \sqrt{\sum_{i=1}^{l_v} c_{k_i}^2}. \tag{6}$$

The QIM for σ_k thus becomes

$$\check{\sigma}_k = \begin{cases} \left\lfloor \frac{\sigma_k}{\Delta_\sigma} + 0.5 \right\rfloor \Delta_\sigma, & \text{if} \quad w_b = 0; \\ \left\lfloor \frac{\sigma_k}{\Delta_\sigma} \right\rfloor \Delta_\sigma + \frac{\Delta_\sigma}{2}, & \text{if} \quad w_b = 1, \end{cases} \tag{7}$$

where Δ_σ is the quantization step for σ_k. The actual modification for the c_{k_i} in $\mathbf{v}_k$ is

$$\check{c}_{k_i} = \frac{\check{\sigma}_k}{\|\mathbf{v}_k + \boldsymbol{\varepsilon}\|}(c_{k_i} + \varepsilon_i), \tag{8}$$

where $\boldsymbol{\varepsilon}$ denotes a vector containing a small-value element ε_i in the ith dimension. The main purpose of utilizing $\boldsymbol{\varepsilon}$ in Eq. (8) is to avoid dividing by zero. Theoretically, if we choose $\Delta_\sigma = \sqrt{l_v}\Delta$ in Eq. (7), the power level of quantization error will be identical to that obtained from Eq. (4).

Based on the discussion in Section 2.1, the maximum tolerable distortion, which is related to the quantization step size, is generally proportional to the signal power. As the signal power varies in different bands, we need to investigate how to group the DCT coefficients into vectors for watermarking so that desirable robustness can be achieved in each band. A pilot experiment has been conducted to analyze the power distribution of the low frequency bands. Fig. 1 depicts the average root-mean-square of the first ten frequency bands, each containing 56 DWT coefficients, drawn from a 10-min audio database along with each vertical bar showing the standard deviation in a specific band. The statistical results indicate that the average power in the first band is approximately 2.2 times of that in the second band and 6.6 times of that in the third band. To maintain the robustness roughly at the same extent, we adopt single-element vectors in the first band and pack two coefficients as a vector in the second band. For the third band, six coefficients are collected as a vector. In other words, l_v is set as 1, 2 and 6 for the first three bands. Furthermore, as revealed by the equation of $a_{tmn}(z)$, lower frequency components tend to have lower signal-to-mask ratios. To allow the robustness to be evenly distributed over the entire band, the vector is formed by selecting the coefficients from both ends toward the middle place. Fig. 2 illustrates the arrangement of the vector grouping for the second band.

Eventually, we embed 48, 24, and 8 bits into the first, second and third frequency bands, respectively. For convenience, the watermarking processes in these three bands

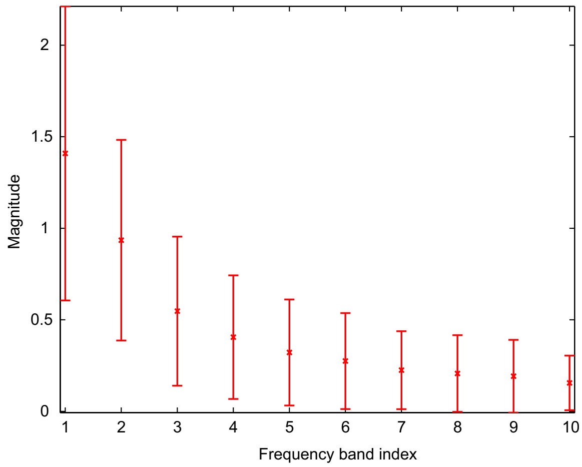


Fig. 1. The means and standard deviations of the root-mean-squares of the DCT coefficients in the first 10 frequency bands. The bar associated each band index is symmetric about the mean with a length of twice the standard deviation.

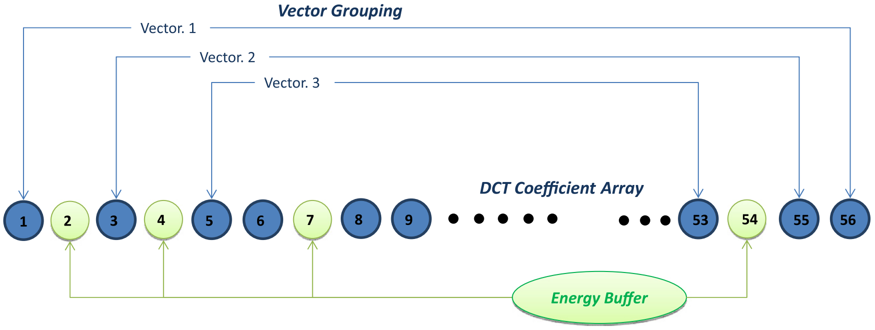
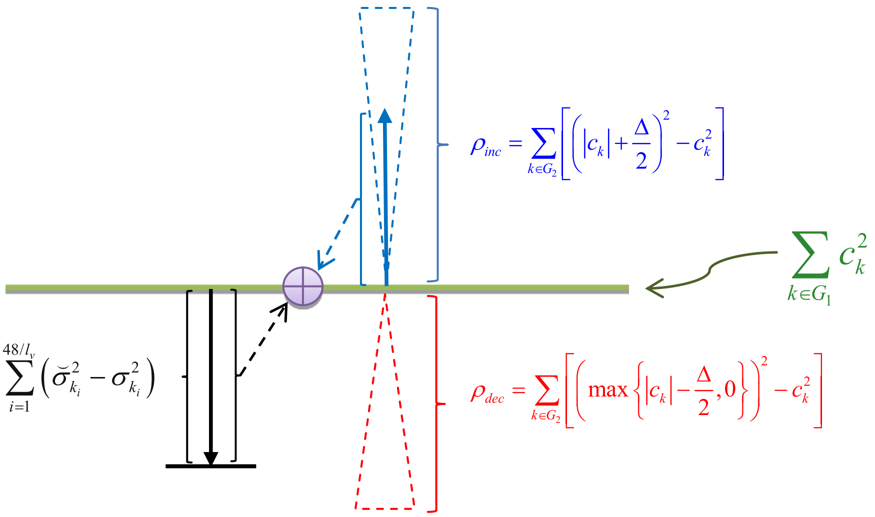


Fig. 2. Demonstration of vector grouping for the DCT coefficients in the 2nd frequency band.



$\rho_{inc}=\sum_{k\in G_2}\left[\left(|c_k|+\frac{\Delta}{2}\right)^2-c_k^2\right]$
$\sum_{k\in G_1} c_k^2$
$\sum_{i=1}^{48/l_v}\left(\breve{\sigma}_{k_i}^2-\sigma_{k_i}^2\right)$
$\rho_{dec}=\sum_{k\in G_2}\left[\left(\max\left\{|c_k|-\frac{\Delta}{2},0\right\}\right)^2-c_k^2\right]$

Fig. 3. Concept of the energy compensation constraint. ρ_{inc} and ρ_{dec} are the upper and lower bounds of the energy variation without leading to perceptual difference.

are denominated as the DCT-b$_1$, DCT-b$_2$ and DCT-b$_3$ respectively. Given that $f_s = 44.1$ kHz and $l_f = l_t + l_w = 4160$, the payload capacities in bits per second (bps) with respect to these three bands are

$$\begin{cases} C_{\text{DCT}-b_1} = 48\frac{f_s}{l_f} = 508.85 \quad \text{(bps)}; \\ C_{\text{DCT}-b_2} = 24\frac{f_s}{l_f} = 254.42 \quad \text{(bps)}; \\ C_{\text{DCT}-b_3} = 8\frac{f_s}{l_f} = 84.81 \quad \text{(bps)}. \end{cases} \tag{9}$$

We note that the DCT is an orthogonal transform with each DCT coefficient denoting the amplitude of a cosine function. The alteration of an arbitrary DCT coefficient does not affect the other coefficients. Consequently, the overall payload capacity, termed C_A, is just the sum of those gathered from the DCT-b$_1$, DCT-b$_2$ and DCT-b$_3$.

$$C_A = C_{\text{DCT}-b_1} + C_{\text{DCT}-b_2} + C_{\text{DCT}-b_3}. \tag{10}$$

2.3. Perceptual constraint

Here we restrict the magnitude change to be less than $\Delta/2$ for each coefficient in G_2. This effect is analogous to the result caused by the QIM. Because of the magnitude restriction, the energy deviation of the c_k's in group G_2 can only vary between ρ_{dec} and ρ_{inc}:

$$\rho_{inc} = \sum_{k\in G_2}\left[\left(|c_k| + \frac{\Delta}{2}\right)^2 - c_k^2\right]; \tag{11}$$

$$\rho_{dec} = \sum_{k\in G_2}\left[\left(\max\left\{|c_k| - \frac{\Delta}{2}, 0\right\}\right)^2 - c_k^2\right]. \tag{12}$$

This implies that the energy variation due to the QIM in G_1 must satisfy the following inequality:

$$-\rho_{inc} \le \sum_{i=1}^{L_{G_1}/l_v}\left(\breve{\sigma}_{k_i}^2 - \sigma_{k_i}^2\right) = \sum_{k\in G_1}\left(\breve{c}_k^2 - c_k^2\right) \le -\rho_{dec}. \tag{13}$$

The energy constraint is visualized in Fig. 3. Note that the QIM shown in Eq. (7) aims at minimizing $\sum_{i=1}^{L_{G_1}/l_v}\left(\breve{\sigma}_{k_i} - \sigma_{k_i}\right)^2$ instead of $\sum_{i=1}^{L_{G_1}/l_v}\left(\breve{\sigma}_{k_i}^2 - \sigma_{k_i}^2\right)$. In case Inequality (13) does not hold, some of the coefficients in G_2 will go beyond their magnitude restrictions in order to compensate the energy variation of the coefficients in G_1. Hence an algorithm is developed in the following to ensure the validity of Inequality (13).

Let us first define a pair of modulated amplitudes

$$\begin{bmatrix}\breve{\sigma}_{k,\{1\}}\\ \breve{\sigma}_{k,\{2\}}\end{bmatrix} = \begin{cases} \begin{bmatrix}\breve{\sigma}_k\\ \breve{\sigma}_k+\Delta_\sigma\end{bmatrix} & \text{if} \quad \breve{\sigma}_k<\sigma_k; \\ \begin{bmatrix}\breve{\sigma}_k\\ \breve{\sigma}_k-\Delta_\sigma\end{bmatrix} & \text{if} \quad \breve{\sigma}_k\ge\sigma_k,\end{cases} \quad \text{for} \quad k=\{1,\ 2,\ \ldots,\ l_v\} \tag{14}$$

where $\breve{\sigma}_{k,\{1\}}$ is the direct outcome of the QIM and $\breve{\sigma}_{k,\{2\}}$ is the suboptimal alternative of the QIM in terms of squared quantization error. The individual energy variation, termed g_k, due to the replacement of $\breve{\sigma}_{k,\{1\}}$ by $\breve{\sigma}_{k,\{2\}}$ for the kth vector is thereby

$$g_k=\breve{\sigma}^2_{k,\{2\}}-\breve{\sigma}^2_{k,\{1\}}. \tag{15}$$

The energy compensation algorithm starts with an initial setup of involved variables:

$$\hat{\sigma}_k=\breve{\sigma}_{k,\{1\}}; \tag{16}$$

$$\eta^{(t)}=\sum_{i=1}^{l_v}\left(\hat{\sigma}^2_{k_i}-\sigma^2_{k_i}\right) \quad \text{with} \quad \eta^{(0)}=\sum_{i=1}^{l_v}\breve{\sigma}^2_{k_i,\{1\}}-\sum_{i=1}^{l_v}\sigma^2_{k_i}. \tag{17}$$

where $\eta^{(t)}$ denotes the energy deviation at the tth iteration. The subsequent part is an iterative procedure consisting of three steps as follows:

Step 1: Within the tth iteration, we end the algorithm whenever $-\rho_{inc}\le\eta^{(t)}\le-\rho_{dec}$. If either $\eta^{(t)}>-\rho_{dec}$ or $-\rho_{inc}>\eta^{(t)}$ occurs, we search for the vector norm σ_K that mostly reduces the energy deviation, i.e.,

$$K=\underset{k}{\operatorname{argmin}}\left|\eta^{(t)}+g_k\right|. \tag{18}$$

Step 2: A new energy deviation is subsequently obtained by

$$\eta^{(t+1)}=\eta^{(t)}+g_K. \tag{19}$$

Step 3: The value of $\eta^{(t+1)}$ is examined. If $|\eta^{(t+1)}|<|\eta^{(t)}|$, we assign $\hat{\sigma}_K=\breve{\sigma}_{K,\{2\}}$ and $g_K=\infty$ before returning to Step 1. Otherwise, the algorithm is terminated.

When performing the QIM in Eqs. (4) or (7), the use of a larger Δ can enhance the watermark robustness but degrade the audio quality. On the other hand, using a smaller Δ avails the imperceptibility but impairs the robustness. Our solution to this dilemma is to raise Δ to the maximum level that is tolerable by the human auditory system. In other words, the amount of noise due to the choice of different Δ must be confined below the auditory masking threshold. Apparently, establishing a link between E_{mask} and Δ is of paramount importance.

Owing to the formulation shown in Eq. (14), the difference between $\hat{\sigma}_k$ and σ_k presumably exhibits a uniform distribution over $[-\Delta/2,\ \Delta/2]$ if $\hat{\sigma}_k$ is chosen as $\breve{\sigma}_{k,\{1\}}$ and a uniform distribution over $[-\Delta,\ -\Delta/2]\cup[\Delta/2,\ \Delta]$ if $\hat{\sigma}_k$ is chosen as $\breve{\sigma}_{k,\{2\}}$. In the worst scenario where all the modified coefficients deviate from their original values by $\Delta/2$ in G_2, the overall energy variance E_{var} thus becomes

$$\begin{aligned} E_{var} &= L_{G_1}/l_v\times E\left[\sum_{i=1}^{l_v}\left(\breve{\sigma}_{k_i}-\sigma_{k_i}\right)^2\right]+L_{G_2}\times\left(\frac{\Delta}{2}\right)^2 \\ &=48\left((1-p)\frac{\Delta^2}{12}+p\frac{7\Delta^2}{12}\right)+8\frac{\Delta^2}{4}, \end{aligned} \tag{20}$$

where $E[\bullet]$ denotes the expectation of energy variation drawn from G_1. The variable p is the probability of choosing $\breve{\sigma}_{k,\{2\}}$ instead of $\breve{\sigma}_{k,\{1\}}$. Based on our experimental observations, p may vary from a small value of 0.010 in the first band to a large value of 0.169 in the third band for various audio signals. For simplicity, its value is assigned as 0.1 in this study. By letting E_{var} equal E_{mask}, we have

$$8.4\Delta^2=10^{\frac{a_{tmn}(z_{rep})}{10}}\times E_c, \tag{21}$$

or equivalently,

$$\Delta_\sigma=\sqrt{l_v}\Delta=\sqrt{l_v}\times\sqrt{10^{\frac{a_{tmn}(z_{rep})}{10}}\times E_c/8.4}. \tag{22}$$

Theoretically, using the above derived Δ_σ will make the embedded watermark imperceptible.

2.4. Start-up of QIM adjustment

The abovementioned energy-compensation algorithm is executed in a sequential manner, i.e., one vector norm after another. By using the proposed iterative algorithm, the inequality condition $-\rho_{inc}\le\eta^{(t)}\le-\rho_{dec}$ is often achieved within a few iterations. However, this algorithm will occasionally run into a stalemate in the third frequency band, where all the coefficients in G_2 have reached their magnitude confinement but still fail to satisfy the energy balance requirement. The reason can be ascribable to the limited available number of involved vectors in G_1 and the bounded capacity of energy absorption in G_2. As the deadlock is often accompanied by a vector of a relatively large magnitude, it appears that the first picked vector is vitally important. A mechanism is devised in the following to resolve the predicament. In a situation where $\eta=\sum_{k=1}^{l_v}\left(\hat{\sigma}^2_k-\sigma^2_k\right)>-\rho_{dec}$, we seek the negative g_k's at first and sort them in descending order, i.e., $0\ge g_{k_1}\ge g_{k_2}\ge\cdots\ge g_{k_n}$, where k_i denotes the position in the original sequence and n complies with the inequality relationship $n\le L_{G_1}/l_v$. Next, we identify the smallest index I that satisfies

$$\eta+\sum_{i=1}^{I}g_{k_i}\le-\rho_{dec} \tag{23}$$

and the smallest J that satisfies

$$\eta+g_{k_J}\le-\rho_{dec}. \tag{24}$$

The occurrence of the condition $(I=J)$ and $(J>1)$ implies that we can simply change one single vector norm $\breve{\sigma}^2_{k_I,\{1\}}$ to $\breve{\sigma}^2_{k_I,\{2\}}$ instead of a batch of $\breve{\sigma}_{k_i,\{2\}}$'s to fulfill the requirement of $\eta+\sum_{i=1}^{I}g_{k_i}\le-\rho_{dec}$.

The procedure of dealing with the situation $\eta<-\rho_{inc}$ is similar to that presented above. We seek the positive g_k's and sort them in ascending order, i.e., $0\le g_{k_1}\le g_{k_2}\le\cdots\le g_{k_p}$. Next, we examine whether the smallest I in $\eta+\sum_{i=1}^{I}g_{k_i}\ge-\rho_{inc}$ and the smallest J in $\eta+g_{k_J}\ge-\rho_{inc}$ are identical. In case $I=J>1$, the vector magnitude $\hat{\sigma}_{k_I}$ is modified as $\breve{\sigma}_{k_I,\{2\}}$. Eventually, the start-up procedure begins with the search of I and J and then verifies the existence of $I=J>1$. Once the initial index is determined, the remaining process for the energy compensation is just like that mentioned in Section 2.3.

2.5. Energy-compensation for the DCT coefficients in the buffer

Following the QIM process executed on the DCT coefficients in G_1, we adjust the coefficient magnitudes in G_2 to counteract the energy deviation. Our strategy here is to evenly distribute the energy deviation $\eta^{(T)}$ to the DCT coefficients in G_2. Here $\eta^{(T)}$ represents the result derived from the iterative algorithm presented in Section 2.3 together with the start-up condition given in Section 2.4. The mode used to process a negative $\eta^{(T)}$ is somewhat different from that with a positive $\eta^{(T)}$. In a situation where $\eta^{(T)}<0$, the amplitude for the mth coefficient is simply modified as

$$\hat{c}_m = \operatorname{sgn}(c_m)\left(c_m^2 - \frac{\eta^{(T)}}{L_{G_2}}\right)^{1/2} \quad \text{for} \quad m \in G_2 \tag{25}$$

where sgn(x) is the sign function defined as

$$\operatorname{sgn}(x) = \begin{cases} 1, & \text{if} \quad x \ge 0; \\ -1, & \text{if} \quad x < 0. \end{cases} \tag{26}$$

When $\eta^{(T)}>0$, every coefficient magnitude in G_2 is supposedly decreased by a certain amount. However, the maximum permissible reduction for each coefficient is limited by its own magnitude. An algorithmic procedure is proposed below to resolve the difficulty. The entire procedure consists of only two steps. First, we sort the coefficient magnitudes in G_2 such that

$$\left|c_{m_1}\right| \le \left|c_{m_2}\right| \le \cdots \le \left|c_{m_{L_{G_2}-1}}\right| \le \left|c_{m_{L_{G_2}}}\right|, \quad m_k \in G_2. \tag{27}$$

Next, we derive the corresponding coefficients with the indexes counting from c_{m_1} to $c_{m_{L_{G_2}}}$:

$$\hat{c}_{m_k} = \operatorname{sgn}\left(c_{m_k}\right)\left(\max\left\{0, c_{m_k}^2 - \frac{\eta_k}{L_{G_2}-k}\right\}\right)^{1/2}, \quad \text{for} \quad k = 1, 2, \ldots, L_{G_2}, \tag{28}$$

where η_k is calculated recursively

$$\eta_{k+1} = \eta_k - \left(c_{m_k}^2 - \hat{c}_{m_k}^2\right) \tag{29}$$

with the initial setup $\eta_1 = \eta^{(T)}$.

The watermarking process for each frequency band ends whenever all the coefficients in G_1 and G_2 are properly modified. Throughout the modifications by either Eqs. (25) or (28), the overall energy for the 56 DCT coefficients in a specific band remains intact, i.e.,

$$\sum_{k \in G_1} \hat{c}_k^2 + \sum_{k \in G_2} \hat{c}_k^2 = E_c = \sum_{k=1}^{56} c_k^2. \tag{30}$$

Consequently, the quantization step Δ derived from $\hat{c}_k$'s is the same as that from c_k's. Such a feature allows us to retrieve Δ directly from the watermarked signal. In other words, there is no need to refer to the original audio signal or side information while extracting the embedded binary bits.

2.6. Smooth transition across frames

The modulation of the DCT coefficients may occasionally lead to an abrupt discontinuity in a frame boundary. A sudden change in an audio signal is often perceived like a click. To avoid such an artefact, a short interval of 64 samples is employed to render a smooth transition across frames. After obtaining the watermarked audio signal $\hat{s}(k)$ located in the watermarking section via the inverse DCT, we define the gaps at the left and right boundaries of the nth frame as α_n and β_n respectively:

$$\alpha_n = \hat{s}\big((n-1)l_f + l_t + 1\big) - s\big((n-1)l_f + l_t + 1\big); \tag{31}$$

$$\beta_n = \hat{s}(nl_f) - s(nl_f). \tag{32}$$

A linear interpolation approach is employed to eliminate the sharp transitions occurring at the end of the previous frame and at the beginning of the current frame.

$$\hat{s}\big((n-1)l_f + k\big) = s\big((n-1)l_f + k\big) + \beta_{n-1} + (\alpha_n - \beta_{n-1})\frac{k}{l_t+1}, \quad \text{for} \quad k = 1, 2, \ldots, l_t \tag{33}$$

where β_0 is assigned as zero. Fig. 4 illustrates the effect due to the employment of Eq. (33). As revealed by the audio waveforms, the sudden changes at frame boundaries are rectified and no click sound has ever been heard.

2.7. Frame synchronization and watermark extraction

To summarize the discussion so far, we depict the entire embedding process in Fig. 5. Like many other implementations, the proposed scheme has been equipped with a synchronization technique [14,30] to withstand the desynchronization attacks. The procedure for extracting watermark bits from a watermarked audio is rather simple. Prior to watermark extraction, we identify the position where the

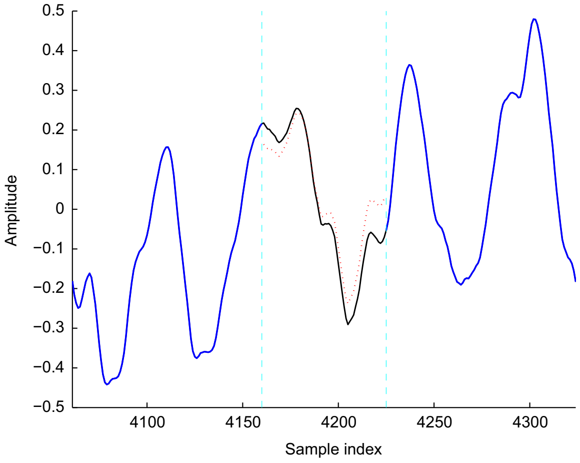


Fig. 4. Illustration of the smoothing across frames. The transition area is located in between two cyan dashed lines. The original signal (as signified by the red dotted line) is modified as the black solid one to render a smooth transition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

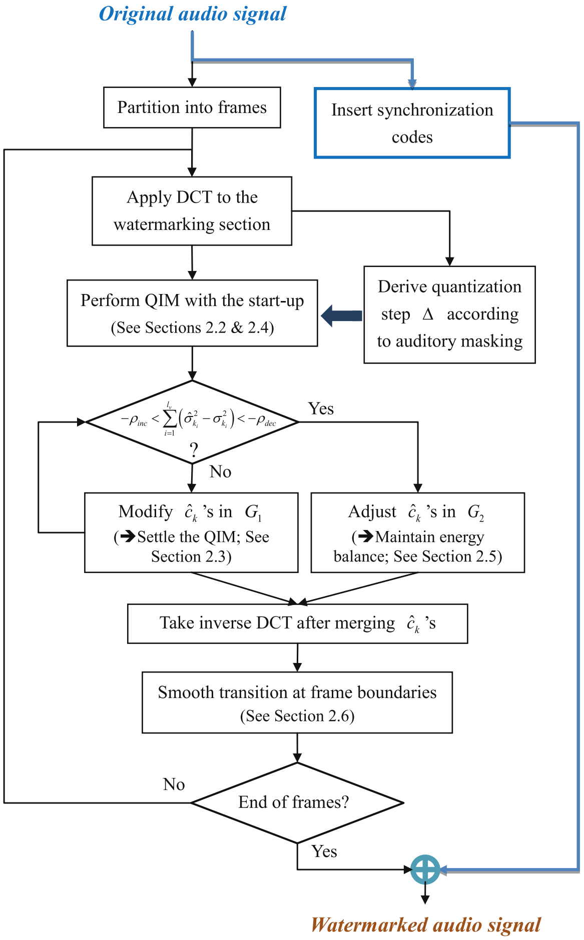

Original audio signal
Partition into frames
Insert synchronization codes
Apply DCT to the watermarking section
Derive quantization step Δ according to auditory masking
Perform QIM with the start-up (See Sections 2.2 & 2.4)
$-\rho_{inc} < \sum_{i=1}^{l_v}\left(\hat{\sigma}_{k_i}^2 - \sigma_{k_i}^2\right) < -\rho_{dec}$?
Yes
No
Modify $\hat{c}_k$'s in G_1 (➔Settle the QIM; See Section 2.3)
Adjust $\hat{c}_k$'s in G_2 (➔Maintain energy balance; See Section 2.5)
Take inverse DCT after merging $\hat{c}_k$'s
Smooth transition at frame boundaries (See Section 2.6)
No
End of frames?
Yes
Watermarked audio signal

Fig. 5. The watermark embedding procedure of the proposed scheme.

watermark is embedded using the standard synchronization technology of digital communications. After we obtain the DCT coefficients from the watermarking section in each frame, the vector norm $\tilde{\sigma}_k$ and quantization step $\tilde{\Delta}_\sigma$ are subsequently acquired as the manner in watermark embedding. The bit $\tilde{w}_b$ residing in each designated vector is determined by

$$\tilde{w}_b = \begin{cases} 1, & \text{if } \left|\tilde{\sigma}_k/\tilde{\Delta}_\sigma - \lfloor \tilde{c}_k/\tilde{\Delta}_\sigma \rfloor - 0.5\right| < 0.25, \\ 0, & \text{otherwise;} \end{cases}$$
$$\text{for} \quad k = \{1, 2, \ldots, L_{G_1}/l_v\}. \tag{34}$$

3. Performance evaluation

The proposed scheme was compared in capacity, imperceptibility and robustness with three other methods, which were named in abbreviated form as the DWT-norm [13], DWT–DCT [23] and DWT–VDVM [14]. The DWT-norm, DWT–DCT, and DWT–VDVM share the similarities in that they perform audio watermarking using the QIM in the DWT domain. To enhance the robustness, the DWT-norm and DWT–VDVM apply the QIM to the vectors formed by the DWT coefficients. In this study, the implementation of the DWT-norm followed the specification in [23], which came up with a payload capacity of 102.4 bps. In contrast, the DWT–DCT and DWT–VDVM were carried out on frames of size 4096 with a payload capacity of 602.93 bps. The DWT–VDVM employed a 5th level DWT to decompose the audio signal and embedded 48 and 8 bits respectively into the 5th approximation and detail subbands roughly at the frequency range of $\left[0,\ 0.5f_s/2^5\right]$ and $\left[0.5f_s/2^5,\ 0.5f_s/2^4\right]$ for each frame. As for the DWT–DCT, we performed a 4^{th} level DWT over the entire audio signal and took the DCT of the 4th level approximation and detail DWT coefficients for each frame. A total of 56 bits was implanted in the first 56 DWT–DCT coefficients in the approximation subband in accordance with the QIM rule. As suggested in [23], the quantization step S for the dither QIM was formulated as

$$S = \psi \times \frac{\left\lfloor \left(\overline{|A(i)|} + \overline{|D(i)|}\right) \times 1000 + 0.5 \right\rfloor}{1000}, \tag{35}$$

where $\overline{|A(i)|}$ and $\overline{|D(i)|}$ represent the mean values of the magnitude DWT–DCT coefficients in the 4th level approximation and detail subbands, respectively. ψ was tentatively chosen as 0.55 to reach a satisfactory tradeoff between robustness and imperceptivity.

The test materials comprised twenty 30-s music clips collected from various CD albums, consisting of vocal arrangements and ensembles of musical instruments. These music clips can be classified into 5 categories including 6 classic music pieces, 4 pop music pieces, 4 rock music pieces, 5 soundtracks, and 1 voice narration with easy listening background music. All audio signals were sampled at 44.1 kHz at 16-bit resolution. The watermark bits for the test were a series of alternate 1's and 0's long enough to cover the entire host signal. Such an arrangement is particularly useful when we want to perform a fair comparison for watermarking methods with different capacities.

The quality of the watermarked audio signal with N samples is evaluated using the signal-to-noise ratio (SNR) defined as in Eq. (36) along with the perceptual evaluation of audio quality (PEAQ) [31].

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{n=1}^{N} \check{s}^2(n)}{\sum_{n=1}^{N} \left(\check{s}(n) - s(n)\right)^2}\right). \tag{36}$$

The PEAQ renders an objective difference grade (ODG) between -4 and 0, signifying a perceptual impression from “very annoying” to “imperceptible”.

Table 1 presents the average ODG's and SNR's for the compared schemes. Basically, the ODG's obtained from all the methods considered in this study are very close to or sometimes even slightly greater than zero, implying that the embedded watermarks are virtually transparent. Because the final result is derived from an artificial neural network that simulates the human auditory system, the PEAQ may come up with a value higher than 0. The data shown in Table 1 clearly exhibit the merits of the proposed scheme. When the watermark bits are solely embedded in a single band, the average ODG's are above zero and the

Table 1
Statistics of the measured SNR's and ODG's along with the payload capacities. The data in the second and third columns are interpreted as "mean [± standard deviation]".

Watermarking schemes	SNR	ODG	Payload (bps)
DWT-norm	24.629 [± 2.253]	−0.305 [± 0.565]	102.4
DWT–DCT	20.893 [± 1.038]	−0.040 [± 0.211]	602.93
DWT–VDVM	20.803 [± 0.327]	−0.127 [± 0.109]	602.93
DCT-b$_1$	20.381 [± 2.366]	0.049 [± 0.070]	508.85
DCT-b$_2$	23.366 [± 1.936]	0.090 [± 0.036]	254.42
DCT-b$_3$	26.436 [± 2.719]	0.050 [± 0.100]	84.81
DCT-b$_A$	17.512 [± 0.521]	−0.112 [± 0.097]	848.08

Table 2
Average bit error rates (in %) for the compared watermarking schemes.

Attack type	DWT-norm	DWT–DCT	DWT–VDVM	DCT-b$_1$	DCT-b$_2$	DCT-b$_3$	DCT-b$_A$
0. None	0.00	0.11	0.00	0.00	0.00	0.00	0.00
A	0.00	0.12	0.00	0.00	0.00	0.00	0.00
B	0.00	0.16	0.00	0.03	0.02	0.00	0.03
C	76.73	0.11	0.00	0.00	0.00	0.00	0.00
D	0.00	0.40	0.02	0.21	0.06	0.09	0.15
E	0.00	1.93	0.55	1.07	0.73	2.88	1.15
F	0.82	0.11	0.00	0.00	0.00	0.00	0.00
G	49.14	47.31	41.41	3.42	35.71	50.18	17.81
H	46.97	0.65	0.13	0.43	0.15	0.69	0.37
I	0.93	4.81	1.05	3.96	3.01	0.60	3.35
J	0.00	0.50	0.34	1.07	0.99	2.13	1.16
K	0.00	0.14	0.02	0.01	0.02	0.03	0.01
L	0.99	2.56	3.10	2.72	4.44	5.53	3.56

corresponding standard deviations remain low, indicating that the watermarked audio signals are perceptually indistinguishable from the original ones and the resultant quality is very steady. The average SNR's for the DCT-b$_1$, DCT-b$_2$ and DCT-b$_3$ are 20.381, 23.366 and 26.436 dB respectively. These values reflect the noise intensities caused by watermarking in different bands. The corresponding large standard deviations signify the energy fluctuation across frames as well as frequency bands. However, once the DCT embedding takes effect for all three bands, the resulting standard deviation in the SNR's remarkably decreases. The reason is ascribable to the fact that the signal energy is mostly concentrated in the frequencies below 1 kHz. Maintaining the embedding strength just below the masking threshold in this range thus renders a relatively stable SNR. Owing to the audio masking, the resultant ODG's can still be near zero even at an average SNR as low as 17.51 dB.

To assess the robustness against various attacks, this study examines the bit error rates (BER) between the original watermark W and the recovered watermark $\tilde{W}$:

$$\mathrm{BER}\left(W,\tilde{W}\right)=\frac{\sum_{m=1}^{M} W(m)\oplus\tilde{W}(m)}{M} \tag{37}$$

where $\oplus$ stands for the exclusive-OR operator and M is the length of the watermark bit sequence. The attack types considered in this study are as follows:

(A) Resampling: conducting down-sampling to 22,050 Hz and then up-sampling back to 44,100 Hz.
(B) Requantization: quantizing the watermarked signal to 8 bits/sample and then back to 16 bits/sample.
(C) Amplitude scaling: scaling the amplitude of the watermarked signal by 0.85.
(D) Noise corruption: adding zero-mean white Gaussian noise to the watermarked audio signal with SNR=30 dB.
(E) Noise corruption: adding zero-mean white Gaussian noise to the watermarked audio signal with SNR=20 dB.
(F) Lowpass filtering (I): applying a lowpass filter with a cutoff frequency of 4 kHz.
(G) Lowpass filtering (II): applying a lowpass filter with a cutoff frequency of 500 Hz.
(H) DA/AD conversion: converting the digital audio file to an analog signal and then resampling the analog signal at 44.1 kHz. The DA/AD conversion is performed through an onboard Realtek ALC892 audio codec, of which the line-out is linked with the line-in using a cable line during playback and recording.
(I) Echo addition: adding an echo signal with a delay of 50 ms and a decay to 5% to the watermarked audio signal.
(J) Jittering: randomly deleting or adding one sample for every 100 samples within each frame.
(K) 128 kbps MPEG compression: compressing and decompressing the watermarked audio signal with an MPEG layer III coder at a bit rate of 128 kbps.
(L) 64 kbps MPEG compression: compressing and decompressing the watermarked audio signal with an MPEG layer III coder at a bit rate of 64 kbps.

Table 2 lists the BER's of the retrieved watermark under various attacks. It appears that the DWT-norm exceeds the other schemes in attack types such as noise corruption and jittering but completely fails in lowpass filtering (II) and amplitude scaling. The payload capacity of the DWT-norm is just 102.84 bps, which is fairly low in comparison with the others. According to the tabulated data, the performance of the DCT-b$_A$ is generally comparable to that of the DWT–DCT. The DWT–VDVM appears slightly better than the DWT–DCT and DCT-b$_A$. Nevertheless, the DCT-b$_A$ hold a payload capacity 40% more than that of the DWT–DCT and DWT–VDVM.

For the DCT-b$_1$, DCT-b$_2$ and DCT-b$_3$, the resulting BER's remain at a similar level. The DCT-b$_3$ renders inferior BER's in the cases of noise corruption, MP3 compression and jittering, whereas its robustness against the echo addition is the best among all. Such a tendency can be elucidated from the viewpoint of spectral characteristics. Fig. 6 depicts the signal spectrum and spectral perturbation caused by various attacks. Notice that the difference between the signal and perturbation spectra remains constant in echo addition but decreases as the frequency climbs in the cases of noise corruption, MP3 compression and jittering. For these three cases, the binary bits embedded in the high frequency region will suffer more impairments. Moreover, because the DCT-b$_3$ receives serious perturbation from the 64 kbps MP3 compression, the resulting BER turns out to be the worst.

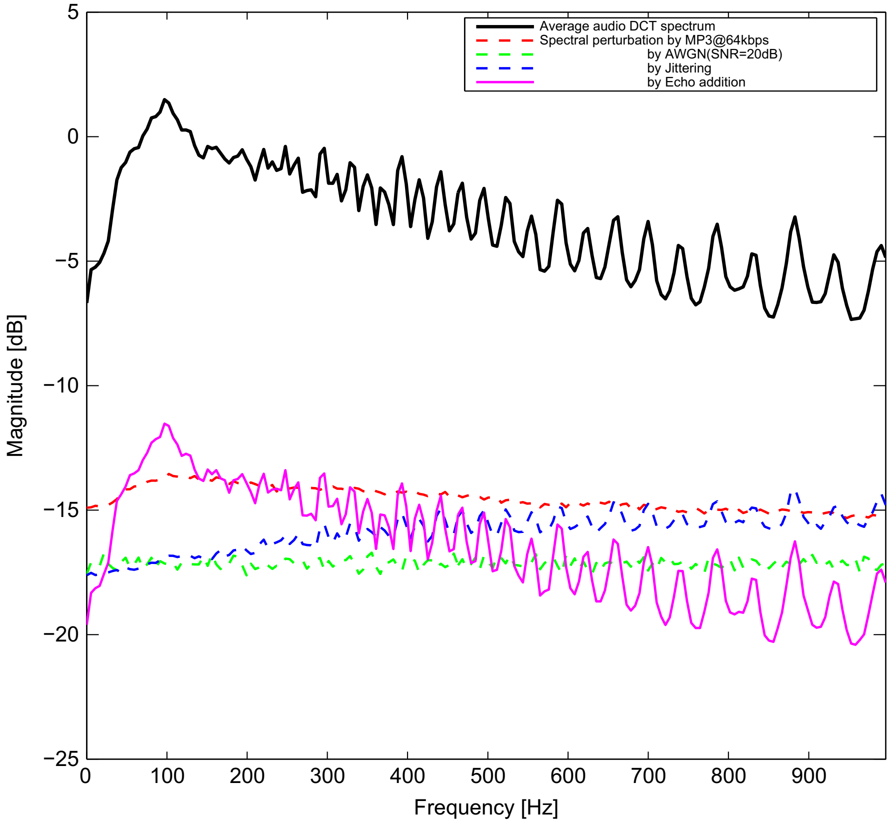


Fig. 6. The average audio spectrum in dB (black solid line) and the spectral perturbations resulting from 64 bps MP3 (red dashed line), noise corruption (green dashed line), jitter (magenta dashed line), and echo addition (pink solid line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

It is worth pointing out that the average BER for the DCT-b_1 in the case of lowpass filtering (II) is just 3.42%, which is significantly less than the others. Such a result is not surprising since the watermarking is performed in a passband below the cutoff frequency. In contrast, the other schemes embed the watermarks over bandwidths spanning across the cutoff frequency. The lowpass filtering can considerably damage the embedded watermark bits.

4. Conclusion

A scheme has been proposed to attain high-performance watermarking by exploiting perceptual masking in the DCT domain. The audio signal is partitioned into non-overlapping frames of length 4160. Within each frame an interval of 64 samples is reserved to smooth the transition across frames and the remaining 4096 samples are converted to the DCT coefficients for efficient watermarking. This study picks three low frequency bands, each consisting of 56 DCT coefficients, to embed an amount of binary information up to 80 bits per frame, which equals to 848.08 bps for audio signals sampled at 44.1 kHz.

During the embedding stage, a perceptual QIM technique is developed to perform robust watermarking without causing perceptible degradation in quality. The embedding strength is adaptively adjusted subject to the auditory masking threshold, which is closely related to the energy variation of the frequency band. To maintain a homologous performance for different frequency bands, more coefficients are grouped as a vector whenever the frequency band energy declines. Moreover, to allow the watermark to be retrievable from the watermarked audio signal, a compensation scheme is proposed to ensure the energy invariability in each band. The PEAQ measure confirms that the watermarked audio is virtually indistinguishable from the original ones even though the SNR is as low as 17.51 dB. Our experimental results also reveal that the proposed DCT-based scheme holds a competent robustness while its payload capacity is much higher than that of the other schemes. Hence it is our conclusion that the proposed scheme can achieve a high performance in all aspects concerned in audio watermarking.

Acknowledgment

This research work was supported by the Ministry of Science and Technology, Taiwan, ROC under Grant MOST 103-2221-E-197-020.

References

[1] P. Bassia, I. Pitas, N. Nikolaidis, Robust audio watermarking in the time domain, IEEE Trans. Multimed. 3 (2) (2001) 232–241.

[2] W.-N. Lie, L.-C. Chang, Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification, IEEE Trans. Multimed. 8 (1) (2006) 46–59.

[3] H. Wang, R. Nishimura, Y. Suzuki, L. Mao, Fuzzy self-adaptive digital audio watermarking based on time-spread echo hiding, Appl. Acoust. 69 (10) (2008) 868–874.

[4] L. Wei, X. Xiangyang, L. Peizhong, Localized audio watermarking technique robust against time-scale modification, IEEE Trans. Multimed. 8 (1) (2006) 60–69.

[5] R. Tachibana, S. Shimizu, S. Kobayashi, T. Nakamura, An audio watermarking method using a two-dimensional pseudo-random array, Signal Process. 82 (10) (2002) 1455–1469.

[6] D. Megías, J. Serra-Ruiz, M. Fallahpour, Efficient self-synchronised blind audio watermarking system based on time domain and FFT amplitude modification, Signal Process. 90 (12) (2010) 3078–3092.

[7] H.-H. Tsai, J.-S. Cheng, P.-T. Yu, Audio watermarking based on HAS and neural networks in DCT domain, EURASIP J. Adv. Signal Process. 2003 (3) (2003) 764030.

[8] X.-Y. Wang, H. Zhao, A. Novel, Synchronization invariant audio watermarking scheme based on DWT and DCT, IEEE Trans. Signal Process. 54 (12) (2006) 4835–4840.

[9] I.-K. Yeo, H.J. Kim, Modified patchwork algorithm: a novel audio watermarking scheme, IEEE Trans. Speech Audio Process. 11 (4) (2003) 381–386.

[10] B.Y. Lei, I.Y. Soon, Z. Li, Blind and robust audio watermarking scheme based on SVD–DCT, Signal Process. 91 (8) (2011) 1973–1984.

[11] X.-Y. Wang, P.-P. Niu, H.-Y. Yang, A robust digital audio watermarking based on statistics characteristics, Pattern Recognit. 42 (11) (2009) 3057–3064.

[12] S. Wu, J. Huang, D. Huang, Y.Q. Shi, Efficiently self-synchronized audio watermarking for assured audio data transmission, IEEE Trans. Broadcast. 51 (1) (2005) 69–76.

[13] X. Wang, P. Wang, P. Zhang, S. Xu, H. Yang, A norm-space, adaptive, and blind audio watermarking algorithm by discrete wavelet transform, Signal Process. 93 (4) (2013) 913–922.

[14] H.-T. Hu, L.-Y. Hsu, H.-H. Chou, Variable-dimensional vector modulation for perceptual-based DWT blind audio watermarking with adjustable payload capacity, Digit. Signal Process. 31 (2014) 115–123.

[15] X. Li, H.H. Yu, Transparent and robust audio data hiding in cepstrum domain, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2000, pp. 397–400.

[16] S.C. Liu, S.D. Lin, BCH code-based robust audio watermarking in cepstrum domain, J. Inf. Sci. Eng. 22 (3) (2006) 535–543.

[17] H.-T. Hu, W.-H. Chen, A dual cepstrum-based watermarking scheme with self-synchronization, Signal Process. 92 (4) (2012) 1109–1116.

[18] V. Bhat K, I. Sengupta, A. Das, An adaptive audio watermarking based on the singular value decomposition in the wavelet domain, Digit. Signal Process. 20 (6) (2010) 1547–1558.

[19] B. Lei, I.Y. Soon, F. Zhou, Z. Li, H. Lei, A robust audio watermarking scheme based on lifting wavelet transform and singular value decomposition, Signal Process. 92 (9) (2012) 1985–2001.

[20] H.-T. Hu, H.-H. Chou, C. Yu, L.-Y. Hsu, Incorporation of perceptually adaptive QIM with singular value decomposition for blind audio watermarking, EURASIP J. Adv. Signal Process. 1 (2014) (2014) 1–12.

[21] S. Katzenbeisser, F.A.P. Petitcolas, Information Hiding Techniques for Steganography and Digital Watermarking, in: Stefan Katzenbeisser, Fabien A.P. Petitcolas (Eds.), Artech House, Boston, 2000.

[22] B. Chen, G.W. Wornell, Quantization index modulation: a class of provably good methods for digital watermarking and information embedding, IEEE Trans. Inf. Theory 47 (4) (2001) 1423–1443.

[23] X. Wang, W. Qi, P. Niu, A new adaptive digital audio watermarking based on support vector regression, IEEE Trans. Audio Speech Lang. Process. 15 (8) (2007) 2270–2277.

[24] H.-T. Hu, L.-Y. Hsu, H.-H. Chou, Perceptual-based DWPT–DCT framework for selective blind audio watermarking, Signal Process. 105 (2014) 316–327.

[25] S.A. Gelfand, Hearing: An Introduction to Psychological and Physiological Acoustics, 4th ed. Marcel Dekker, New York, 2004.

[26] B.C.J. Moore, An Introduction to the Psychology of Hearing, 6th ed. Brill, Leiden, 2013.

[27] X. He, M.S. Scordilis, An enhanced psychoacoustic model based on the discrete wavelet packet transform, J. Frankl. Inst. 343 (7) (2006) 738–755.

[28] T. Painter, A. Spanias, Perceptual coding of digital audio, Proc. IEEE 88 (4) (2000) 451–515.

[29] X. He, Watermarking in Audio: Key Techniques and Technologies, Cambria Press, Youngstown, NY, 2008.

[30] H.-T. Hu, C. Yu, A perceptually adaptive QIM scheme for efficient watermark synchronization, IEICE Trans. Inf. Syst. E95-D (12) (2012) 3097–3100.

[31] P. Kabal, An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality, TSP Lab Technical Report, Department of Electrical & Computer Engineering, McGill University, 2002.